The Assessment of Research Quality in UK Universities: Peer Review or Metrics?[1]

Jim Taylor
Department of Economics
Management School
Lancaster University
Lancaster LA1 4YX
Email: jim.taylor@lancaster.ac.uk

February 2010

Abstract

This paper investigates the extent to which the outcomes of the 2008 Research Assessment Exercise, determined by peer review, can be explained by a set of quantitative indicators, some of which were made available to the review panels. Three cognate units of assessment are examined in detail: business & management, economics & econometrics, and accounting & finance. The central focus is the extent to which the quality of research output, as determined by the RAE panel, can be explained by the quality of journals in which research is mainly published. The main finding is that although a high proportion of the variation between departments in their RAE outcomes can be explained by quantitative indicators, there is insufficient evidence to support the claim that journal quality is a sufficiently accurate predictor of research quality to justify a predominant role in the research assessment process. A further finding is that the judgment of the RAE panels was biased in favour of Russell Group universities. There is also evidence that the decisions of the economics & econometrics panel were biased.

[1] I am extremely grateful to David Stott for matching the ABS journal quality scores to the publications submitted to the RAE. I am also grateful to three referees and to Ian Walker, David Peel, David Otley, Mike Pidd, David Collinson, Robert Fildes and Colin Wren for very helpful and constructive comments on an earlier draft. The author alone is responsible for all views, errors and omissions.

The Assessment of Research Quality in UK Universities: Peer Review or Metrics?

1 INTRODUCTION

There has been intense debate for over two decades about how best to measure the quality of research produced by the university sector. This debate has been particularly vocal in the UK because of a sequence of Research Assessment Exercises (RAEs) which have been used to allocate block research grants to individual universities.[2] Peer assessment has therefore been the order of the day in the UK's higher education sector since the late 1980s. Interest in the UK's attempts to measure research quality has been widespread and its intermittent RAEs have become a 'benchmark' for the development of similar research assessment systems in other countries, particularly Ireland, Australia and New Zealand (Tunzelmann and Mbula, 2003).

Responsibility for measuring research quality in the UK has been delegated to its higher education funding councils, which in turn have appointed subject experts to judge the quality of research submitted for assessment by the universities.[3] Thus, in the 2008 RAE sixty-seven expert subject panels were responsible for assessing up to four outputs from over 68,000 individual researchers[4] in order to rate university departments according to a predetermined research quality scale with four distinct categories, ranging from 'nationally recognised' up to 'world leading'.

Although the merits or otherwise of the peer review system have been extensively discussed within the academic community for as long as the RAE has existed, it was not until the intervention of the Science and Technology Committee of the House of Commons that the funding bodies were forced to seriously consider whether sole reliance on peer review was appropriate. The Science and Technology Committee not only accused the funding councils of shying away from radical change in its assessment of research quality, but also came down heavily in favour of using 'metrics' to support, *or even replace*, the peer review process (House of Commons, 2004a, p.3). The Committee subsequently advocated "a more radical approach, employing a range of metrics to reduce the bureaucratic burden on universities" (House of Commons, 2004b, p.5). It was this recommendation by the Science and Technology Committee that led directly to the UK Government asserting its preference for a metrics-based system of assessment in future appraisals of research by the funding councils (HM Treasury, 2006a).

---

[2] Six RAEs have been undertaken since 1986. The other five were in 1989, 1992, 1996, 2001 and 2008. The 1986 exercise was not used to determine the allocation of research funding but was simply a pilot to develop the process of assessing research quality.

[3] There are separate funding agencies for England, Scotland, Wales and Northern Ireland, but the Higher Education Funding Council for England (HEFCE) plays the lead role in the RAE.

[4] There were 55,015 Category A researchers (with employment contracts) and 13,548 other staff (including retired and semi-retired staff).

The assertion by politicians that ‘metrics’ should play a much bigger role in the research assessment process has led the UK funding councils to take bibliometric data more seriously. The metrics v peer review debate is now intense and answers to questions about the extent to which reliance should be placed on bibliometric data in the assessment of academic research are urgently needed. The primary aim of this paper is therefore to investigate the extent to which the 2008 RAE outcome in three cognate units of assessment can be explained by quantitative indicators related to research activity. The three cognate units are: business & management, economics & econometrics, and accounting & finance. Regression methods are used to evaluate the relationship between: (i) the outcomes determined by peer review in these three units of assessment, and (ii) a set of quantitative indicators. These estimated relationships are then used to assess the potential value of the quantitative indicators for measuring the research quality of individual institutions within these three units of assessment.

The four main findings of this paper are as follows. First, a substantial proportion of the variation (between departments) in the RAE panel’s assessment of research quality can be explained (statistically) by a set of quantitative indicators, including a journal quality index based on the Association of Business Schools’ *Journal Quality Guide*. Second, although journal quality is the main explanatory variable, the explanatory power of the regressions is significantly enhanced by the inclusion of departmental size as an explanatory variable as well as by the influence of a small number of outlying observations. This paper therefore challenges the view that research quality can be accurately estimated from journal quality ratings alone. Third, there is evidence of bias in the judgment of both the business & management panel and the econometrics & economics panel in favour of the Russell Group universities, possibly due to a ‘halo effect’. Fourth, the judgments reached by the economics & econometrics panel appear to have been biased upwards compared to the business & management panel.

The remainder of this paper is organized as follows. Section II discusses the policy background to the debate over the use of bibliometric data in the UK’s research assessment process. Section III reviews recent literature on the potential usefulness of bibliometric data to measure the quality of research output. Section IV outlines a model of the research assessment process in order to specify the variables expected to influence the decisions reached by the RAE panel. Section V presents the results of an empirical analysis of the research ratings awarded by the panels. The implications of these results for the use of bibliometric data in the research assessment process are discussed in section VI. Section VII concludes that although there is still no alternative to peer review in assessing the quality of research, there is nevertheless a clear role for using journal ratings as part of the assessment process.

II POLICY BACKGROUND TO RESEARCH ASSESSMENT IN THE UK

The UK Government has asserted that the assessment of university research “should be simple and cost-effective” (HM Treasury, 2006a: p.31). In considering the cost-cutting options available for reducing the need for a peer review process, the Government’s clear preference has been to identify one or more metrics that could be used to assess research quality. It goes further by identifying research income, citations and research student numbers as examples of what might be appropriate metrics. Indeed, the Treasury report states that "the close correlation between Research Council income and QR income [i.e. block research grants from the funding councils] may provide an opportunity for allocating QR using a radically simpler system" (p.31).The report concludes that "... after 20 years of relying on the RAE to allocate these [QR] funds the Government thinks there is now sufficient evidence to support moving towards a simpler and less burdensome system of allocation" based on 'metrics' (HM Treasury, 2006a: p.33). In 2006, the Government made it clear that following the completion of the 2008 RAE, "the system for assessing research quality and allocating QR funding will be mainly metrics-based" (HM Treasury, 2006a: p.30). On consulting with the higher education funding councils, however, the Government recognized that the appropriateness of metrics for measuring research assessment differed across disciplines. A bibliometric index to assess research was deemed to be appropriate for science, engineering and technology, but for other disciplines "a significantly reduced, light-touch peer review process informed by a range of discipline-specific indicators" should be adopted (HM Treasury, 2006b: p.57). [5]

The funding bodies themselves have, in the meantime, been proactive in investigating the potential value of using metrics more directly. Following the 2001 RAE, the Higher Education Funding Council for England (HEFCE) commissioned a review of the RAE process. The Roberts Report (2003) considered whether there was scope for assessing research performance based on metrics, such as research income, research students and bibliometric data, as an alternative to the complex and labour-intensive peer review process. Replacing the peer review process with one based upon metrics was firmly rejected by Roberts, who came to the conclusion that "the only system which will enjoy both the confidence and the consent of the academic community is one based ultimately upon expert review" (Roberts, 2003, p.7). This was qualified, however, in the very first recommendation of the Roberts Report, which proposed that: "Any system of research assessment designed to identify the best research must be based upon the judgment of experts, who may, if they choose, employ performance indicators to inform their judgment" (Roberts, 2003, p.7). This view is generally supported by the academic community, which has expressed a preference that metrics should play a greater role in supporting, but not replacing, the work of expert reviewers (Roberts, 2003, p.82). The funding bodies' response to the Roberts' proposal, however, was somewhat more guarded. The use of discipline specific metrics was supported "provided it does not take the place of *or unduly influence the judgment* of experts". (HEFCE, 2003, p.1, own italics)

In view of the Government's insistence that metrics should play a bigger part in the research review process, and following the completion of the 2008 RAE, HEFCE began a review of how bibliometric data based on citation counts of publications might be used in future research assessments. Clarifying how bibliometric data will be used in the forthcoming Research Excellence Framework (REF) needs to be done well in advance of the start of the process so that universities know how their research output in the current period will be assessed. After considering several different research performance indicators based on citations, HEFCE's preliminary view was that "bibliometrics are not sufficiently mature to be used

[5] Scepticism over the use of metric-based systems is not confined to the arts and social sciences. It is also evident in the sciences (Hobbs and Stewart 2006).

formulaically or to replace expert review, but there is considerable scope for citation indicators to inform expert review in the REF". The review also recommends that citation indicators should be used only for "selected papers by the staff in each submission, rather than attempt to capture all papers" (HEFCE, 2009, p.2/3).

It therefore seems certain that the research assessment process will take bibliometric data on board in future assessments, though the extent to which such data are used is expected to vary considerably between subjects. We focus here on the potential use of bibliometric data in business & management, economics & econometrics and accounting & finance.

III LITERATURE REVIEW OF METRICS V PEER REVIEW

Although metrics and peer review are clearly not mutually exclusive approaches to assessing the quality of academic research, it is useful to consider these as alternative assessment methods in order to identify their respective strengths and weaknesses. The basic difference between these two approaches is that whereas peer review is intended to be a *direct* measure of the quality of research output based on expert opinion, citations are an *indirect* indicator of research quality.

Methods of assessment based on metrics

The use of bibliometric data to assess the quality of research is attractive because citations measure the number of times a piece of research has been used by, and possibly influenced, the work of other researchers. Furthermore, the availability of citations data through internet sources, such as the Thomson-Reuters *SSCI*, has made it possible to construct citation impact scores at both the individual and departmental level. The purpose of these citation scores is to measure the influence and impact of an individual's or a department's publications.

Measures of research quality based on citations data, however, suffer from a major problem when used to assess recently published research, as in the UK's RAE: the time frame is too narrow due to lengthy publication lags, which can be several years in the social sciences. Even papers published early on in the research assessment cycle may not be cited for several years due to publication lags in the citing papers. There are also marked differences in the speed at which papers are cited following their publication. Nierop (2009), for example, shows that citations of papers in statistics journals take significantly more years to reach their peak than papers in science journals. These problems with citations data *per se* have led to the use of a proxy based on the 'quality' of the journal in which a paper is published. Just as citations of individual articles are used to proxy the quality of research in these articles, the quality of a journal can be used to proxy the quality of all articles published in that journal. Publishing in highly cited journals, for example, has long been important in determining the career paths of academics, especially where tenure track systems are in place (Baruch and Hall, 2004).

Journal quality can be approximated in various ways. The most basic approach is to determine the reputation of journals through a survey of academics within a discipline in order to obtain their views about the relative status of the journals with which they are familiar. This has been done, for example, by McLean *et al*. (2009) for 92 journals in politics and international studies. A major weakness of this approach is that it is highly susceptible to sampling response bias, perhaps even due to organized responses by sub-groups within disciplines (Russell 2009).

A more common method of rating journals is to use a journal impact factor based on the number of citations to articles published in each journal. Such an index has been constructed for economics by Kodrzycki and Yu (2006) from the Thomson-Reuters *Journal Citation Reports* (see also Liebowitz and Palmer, 1984; Laband and Piette, 1994; Liner, 2002; Kalaitzidakis *et al*., 2003). Journal impact factors are widely accepted as being a useful 'broad brush' guide to the quality of journals, though not necessarily to individual articles within journals. The use of journal ratings has the added advantage, in the context of the RAE, that the quality of papers not yet published but which have been accepted for publication can also be proxied.

Journal ratings can also be obtained from the revealed preference of RAE panels in previous rounds of assessment. Geary *et al.* (2004), for example, estimate the quality of journals implied by the decisions made by the RAE business and management panel in the 2001 exercise. They calculate a journal rating score for 526 core business and management journals included in the 2001 RAE. Easton and Easton (2003) undertake a similar exercise for marketing journals. A more ambitious exercise is undertaken by Mingers *et al*. (2009) for the 2008 RAE. They use linear programming techniques to transform the research profiles determined by the RAE panel for business schools into corresponding research profiles for 700 journals in which submitted articles were published.

A further approach is to use a combination of methods. The UK's Association of Business Schools (ABS) has constructed and published a *Journal Quality Guide* since "[t]hose who fund research and evaluate the outcomes need a guide to the quality of the academic outlets in which it is published" (Kelly *et al.* 2009a). The ABS *Guide* is based partly on journal citations and partly on the editorial judgments made by an invited panel of experts in response to feedback from the business school community (Kelly *et al*., 2009a, p. i). According to the ABS, its *Journal Quality Guide* has two major advantages: first, its coverage is considerably wider than journal ratings based solely on citation counts (since many academic journals are not present in any of the published citation lists); second, the *Journal Quality Guide* is subjected to ongoing peer review by a nominated panel of experts.

Some problems with using journal rating scores to measure research quality

The use of journal rating scores to assess research quality has been heavily criticised on several counts, especially when used for assessing the quality of research at the level of the individual researcher. MacDonald and Kam (2007) make a scathing attack on the games played by academics in efforts to establish their research prowess by publishing in top-rated journals. Fleet *et al*. (2000) argue that journal ratings can lead to adverse long-term consequences, both for the individual researcher and for institutions, if used as a basis for staff appraisal and promotions, especially if the ratings are locally constructed and simply reflect the subjective views of existing staff members. Lee (2007) goes further. He argues that the definition of research quality adopted by the RAE has been driven by mainstream (neo-classical) economists who publish in a narrow list of highly cited economics journals which in turn publish only articles based on the neo-classical paradigm. According to Lee, orthodox economists have used the RAE to reward departments committed to mainstream economics while simultaneously downgrading the value placed on heterodox economics. Ozbilgin (2009) and Nkomo (2009) make similar claims for business and management journals, arguing that the top-rated journals cater primarily for the US market and discriminate against ‘outsiders’.

Journal ratings have also been criticised for being too narrow (Adler and Harzing, 2009). By definition, journal ratings ignore citations of books, book chapters, conference presentations, working papers and internet publications, all of which may have a substantial impact, both inside and outside of academia. Adler and Harzing argue that the use of journal ratings in assessing the value of research could seriously harm the type and usefulness of research that is being undertaken. Books and chapters in books, for example, allow researchers to be more innovative and controversial than is often possible within the tight confines of established methodology adopted by journals through the refereeing process. Researchers aiming for publication in journals may feel constrained to take a more traditional approach in order to reduce the probability of rejection.[6]

Adler and Harzing consequently argue in favour of a more comprehensive approach to measuring research output and its impact. They propose the adoption of a wide range of research indicators so that the narrow focus on journal publications can be counter-balanced by citations of non-journal publications, which are captured by internet-based services such as *Google Scholar* and *Publish or Perish*. Both services are now well established in the university sector and their use in measuring research output would help to capture a wider range of impacts than is possible by relying entirely on journal citations. Adler and Harzing would consequently “applaud a proliferation of rankings using varied assessment measures” (p. 91). These could also include citation sources for individual subjects, such as *Research Papers in Economics* (*RePEc*).

Perhaps the most critical problem with all journal quality indices is that they aim to measure the *average* quality of the papers published in each journal. Not all articles published in top journals, however, are of ‘high quality’, at least as indicated by their subsequent citations record. The converse is also the case: some articles in journals of lower status are highly cited (Bessant *et al*. 2000, Singh 2007). Singh, for example, finds that around one-half of articles published in top management journals (using a seven-year citations window) are not in the top tier of cited articles. Furthermore, around one-half of all top-cited articles are not published in ‘top’ journals. A similar picture emerges in medical sciences: Weale *et al*. (2004) find that around 16% of articles published in immunology and surgery journals accounted for over 50% of total citations. Very similar results are obtained by Wall (2009) for economics journals, Seglen (1997) for biochemistry journals and Walter *et al*. (2003) for psychiatry journals. The use of journal impact factors to measure the quality of individual articles therefore “enables research that has made no detectable

[6] This view is rejected by Clarke and Wright (2007), who argue that journals develop and change in response to innovative approaches.

impression on the academic community to steal prestige from more conspicuous articles that happen to appear in the same journal". (Williams 2007, p.568)

Although journal quality ratings are unlikely to be a good guide to the quality of individual articles, they are likely to be more reliable at departmental or faculty level due to the cancelling of errors (Starbuck, 2005). The use of journal quality ratings, however, has so far been rejected by funding bodies in the UK. This is not surprising, given the attitude of RAE panels. Commenting on the assessment of the quality of publications in the 1996 RAE, Cooper and Otley (1998) confirmed that the business and management panel relied heavily on the individual judgment of panel members in reading the submitted research. Although recognising that the perceived quality of journals can provide useful additional information, they argued that journal ratings are unreliable and should not replace the peer review process (see also Otley, 2002; Bessant, 2003; and Paul, 2008).

In the 2008 RAE, the business and management panel decided not to use journal quality ratings *at any stage* in their assessment of publications (HEFCE, 2009b). This was also the case for cognate panels in the same main group (i.e. economics & econometrics and accounting & finance).[7] It is clearly the case, however, that panel members did not have the time to read all submitted outputs in detail, given that this would have required each member to read around 700 outputs in business & management and 250 in economics & econometrics. It therefore seems very likely that the assessment of individual research outputs must have been influenced by place of publication, or alternatively by the researcher's place of work.

Peer review

The peer review method of assessing research quality has not, however, escaped criticism. Bence and Oppenheim (2004) argue that panel members are unlikely to possess the range of expertise that is available to journal editors when selecting referees. Review panels do not have super-human powers. There is no way that a small panel of experts can match the wide range of expertise available to journal editors. This leads Bence and Oppenheim to question the reliability of peer review. They argue that the review process should rely more on metrics and less on the subjective judgment of RAE panels.

The central problem with relying exclusively on peer review is that panel members may be unable to detach their predilections of the quality of research previously produced by individual departments from their independent assessment of the research output submitted to the RAE. Despite considerable, and sometimes bitter, disputes about the role and influence of implicit bias in organizational decision-making, there is a large body of empirical evidence in the social and cognitive psychology literature supporting its existence and pervasiveness (Jost *et al*., 2009; Tetlock and Mitchell, 2009). In the context of the RAE, implicit bias may occur because panel members may be influenced by the 'halo effect' surrounding departments with long-standing records of high quality research (Nisbett and Wilson, 1977; Wilson *et al*., 1993; Reich *et al*., 2007).

---

[7] The chair of the main panel stated that: "The sub-panels assessed virtually all the submitted work by examining it, and did not use its place of publication as an evaluative criterion." (HEFCE, 2009b)

The consequence of this halo effect is that it may significantly bias the judgment of panel members despite their determination to provide accurate and unprejudiced assessments of the submitted research output. Eliminating such bias is impossible in the RAE context since the source of all submitted research output is known. Anonymity is not possible no matter how hard panel members may (but probably do not even) try to operate a blind assessment system. Indeed, the requirement to assess a department's esteem, alongside its research output, means that halo effects are an explicit part of the evaluation criteria.

Quantitative studies of research assessment in previous RAEs

Several quantitative studies have been undertaken, dating back to the 1989 RAE, relating measurable research inputs to outcomes using data from consecutive RAEs. Johnes *et al*., (1993), for example, show that differences between institutions in the RAE outcome within major subject groups are related to factors such as departmental size, the research rating of other departments in the same institution (an institution level halo effect), and research income per staff member. More detailed statistical analyses became possible for the 1992 RAE due to the collection and publication of data relating to the quantity and type of research output (at subject group level). This included, for example, the number of articles published in refereed journals as well as data on Research Council grants (see Taylor, 1994, 1995, 1996). Analyses based on the 1992 RAE show that a high proportion of the variation in research ratings between institutions within subject groups is accounted for by only a handful of variables, the prime ones being departmental size, research papers published in refereed journals and research income, especially 'blue chip' income from the research councils.

These previous investigations of the statistical relationship between research inputs and outcomes stop short of suggesting that quantitative data may be valuable in the research assessment process, although this is clearly one of the reasons such data are collected and made available to the RAE panels. A more proactive line is taken by Doyle *et al*. (1995), who argue that data collected for the RAE could be used much more effectively by the business and management panel to inform its judgments. Using data collected for the 1992 RAE, Doyle *et al*. show that 87% of the variation in decisions reached by the business and management panel can be accounted for by a set of nine variables, all of which are statistically significant at 5% or less.[8] Having identified the factors that 'explain', in a statistical sense, the decisions reached by the panel, Doyle *et al*. then demonstrate how the data can be used to assess the research performance of individual units of assessment based on linear programming techniques. They argue that providing such information would allow the panel to focus on deciding what special factors needed to be considered to adjust the initial quantitative estimate of each unit's research performance in reaching its own final decision on research quality.

Similar studies have been undertaken in other social science disciplines. Butler and McAllister (2009) argue that the 2001 RAE outcomes for political science can be accurately predicted from a combination of citations data (computed for each submitted publication) and other quantitative indicators

[8] These are as follows: number of staff submitted, number of publications (journal articles, popular articles and book reviews), Research Council income, research studentships, and three binary variables (old/new university, panel member/or not, and located in England/other UK country).

such as departmental size. On the basis of their findings, they propose the development of a set of robust quantitative indicators in order to obviate the need for an extensive peer review process. They further suggest that "only a small panel of experts would be needed to scrutinize the anomalies and to weigh up the merits of any claimed disadvantage from the choice of metrics" (p. 15). Johnston (2009) and Russell (2009) reject absolutely Butler's and McAllister's findings. They point not only to the inappropriate use of ordinary least squares regression (rather than multinomial logit) to model categorical data, but also challenge Butler's and McAllister's claim that the predictions are 'very close' to the actual outcomes. Their model fails to accurately predict 30 of the 69 outcomes, which is not surprising given that they use the wrong statistical model and that their estimated equation explains only 62% of the variation in RAE outcomes. It could, of course, be argued that it is not the regression model that got it wrong but the RAE panel. It would not, however, be easy to convince the stakeholders, particularly the researchers themselves, that this was the case.

The most recent attempt to demonstrate that quantitative data can explain a large proportion of the variation in research assessment outcomes is by Kelly *et al*. (2009b) using data from the 2008 RAE. They find that 90% of the variation between institutions in the overall outcome of the 2008 RAE for the business & management group is 'explained' by just four variables: the journal rating obtained from the ABS's *Journal Quality Guide*, the number of staff submitted for assessment, the percentage of research income emanating from the research councils, and the percentage of submitted research output published in journals. Kelly *et al.* conclude that "the *Guide* in large measure mirrors the quality judgements made by the RAE panel"; and further that "simple volume and quality indicators based on journal rankings might offer a sound metric to help inform funding decisions" (p.15-16).

Despite the length and complexity of the debate over the use of quantitative indicators in the research assessment process, there is still no clear view about the role that such indicators should play. Academic researchers, the funding bodies and politicians have all been keen to express their views about the potential role of quantitative indicators such as journal quality indices and citations, but a consensus has still to emerge regarding the way in which such indicators should be used and whether or not they should supplement or replace the peer review process. This choice has been highlighted in the business and management area by the proposal from the ABS that its *Journal Quality Guide* provides a sound basis for informing future funding decisions (Kelly *et al*. 2009b, p.15/16).

IV MODELING THE 2008 RAE OUTCOMES

This section specifies a testable model of the outcomes, at departmental level, of the 2008 RAE and is motivated by the literature survey in section III. It is necessary to note at the outset that the purpose of modeling the RAE outcome is to investigate the extent to which quantitative data relating to each institution's research activities is capable of replicating the research quality decisions reached by the subject panels. The model is therefore 'statistical' and is not meant to imply causality. In other words, the model attempts to explain the actual RAE outcomes and not some theoretical 'true' research quality. The aim is therefore to identify a set of quantitative indicators which capture crucial aspects of the factors that may have influenced the RAE panel in its deliberations. Furthermore, since a department's overall research performance is a weighted average of its research output (70%), esteem (20%) and research environment (10%), it is appropriate to 'explain' the outcomes of each component separately. A simplified schematic illustration of the model is provided in Figure 1.

The outcomes of the 2008 RAE

In the first four RAEs, an expert panel awarded each unit of assessment a single score (e.g. on a 7-point scale in the 2001 RAE) reflecting the quality of its research. This procedure was abandoned in the 2008 RAE and replaced by profiling, whereby all research output was distributed between five classes according to the degree to which the research met certain standards (4*=world-leading, 3*=internationally excellent, 2*=international, 1*=national, 0=unclassified). Moreover, this criterion was applied to all three aspects of research performance: research output, esteem and research environment. An overall research profile was then obtained for each department within each subject group by calculating the weighted average of these three research categories.

As shown in Table 1, the research profiles for the three units of assessment investigated here are very different. Economics & econometrics, for example, has 76% of its researchers in the combined 3* and 4* classes in the overall classification, compared to only 53% for business & management and 44% for accounting & finance. The same disparities are evident in the other three research quality categories. In the esteem indicators category, for example, economics & econometrics classifies 88% of its researchers in the combined 4* and 3* categories compared to 60% in business & management and 49% in accounting and finance.

A more succinct way of expressing the RAE outcomes is provided in Table 2, which compresses the profiles into a single score by weighting each of the four classes and then summing. The simplest weighting system is as follows: 4 = 4*, 3 = 3* 2 = 2*, 1 = 1* and 0 = unclassified (see notes to Table 2). This way of calculating the RAE outcomes has the disadvantage that the 'distance' between each category is linear (e.g. world-leading research output is assumed to be 'worth' four times as much as output ranked at 'national' level). An alternative is to use the weighting system implied by the formula used by HEFCE for distributing research funding[9], which is as follows: 4*=7, 3*=3, 2*=1 and other=0. Since the correlation between the RAE scores obtained by these two methods are very highly correlated (rank correlation=0.992 for the overall RAE score for business & management), the choice of weighting system has virtually no effect on the results of the statistical analysis. An advantage of using linear weights, rather than the weights implied by the HEFCE formula, is that the estimated regression coefficients are easier to interpret. The mean scores given in Table 2 are therefore based on the linear weights. The main point to note about the mean scores in Table 2 is that the score for economics & econometrics is significantly greater (mainly at the

[9] See http://www.hefcw.ac.uk/documents/policy_areas/research/qr%20funding%20method.pdf.

0.1% level) than the mean scores for both business & management and for accounting & finance. We return to this performance differential below.

Research output

We turn first to the main component of the overall outcome, namely research output. Previous research indicates that this is likely to be determined predominantly by the quality of the publications each unit of assessment submits for appraisal to the RAE panel. Since there is no direct measure of the quality of research output, it is necessary to use an indirect measure based on journal quality rankings. The method adopted in the present paper is based on the journal rating exercise undertaken by the ABS (Kelly *et al*., 2009a). The estimated quality of publications is therefore based entirely on the designated quality of the journal in which each article is published using the most up-to-date version of the ABS *Journal Quality Guide* since this is the most complete.

We know from the literature survey that a fundamental disadvantage of using journal rating scores to proxy the quality of individual publications is that not all publications in highly rated journals are of high quality; and conversely, many publications in lower rated journals are of high quality (Bessant *et al*., 2000; Singh, 2007). The only response to this weakness of journal rating scores is that the averaging process (over an entire department's publications) may help to wash out the errors emanating from this source at individual staff level (Starbuck, 2005). Added to this fundamental flaw in all journal rating scores, however, are two further problems. First, the method is applicable only to research published in journals (Adler and Harzing, 2009). There is, as yet, no comparable mechanism for assessing the quality of books, chapters in books or other forms of publication. Second, not all journals are covered by the ABS *Journal Quality Guide*. Out of a total of 12,575 outputs submitted in business & management, for example, an ABS score was not available for 2,267outputs. The fact that around 82% of the research outputs submitted for assessment (in all three subject areas) are covered by the ABS score helps to overcome, to some extent, this problem of incomplete coverage (see Table 3).[10]

But leaving out nearly 20% of the outputs submitted to the RAE may seriously reduce the validity of the statistical analysis. This suggests the need to substitute some plausible values for the missing observations. The two most obvious approaches to filling in the missing values are as follows. First, an arbitrary score of unity can be awarded to replace all missing observations. This may underestimate the research panel's evaluation of 'non-covered' outputs, however, since a greater proportion seem likely to be rated better than 1* by the panel than are likely to be rated as 'unclassified' (hence receiving a zero score). We might therefore expect an implicit score greater than unity to be more plausible (on average) for research outputs not covered by the ABS *Journal Quality Guide*. The second method of replacing missing observations is to assume that all missing outputs are 'worth' the mean score of the outputs for which an ABS score is available. This may overestimate the research panel's evaluation of 'non-covered' outputs, however, especially for second, third or fourth publications which are neither articles nor authored books.

[10] The percentage of submitted outputs covered by the ABS score is 82% for business & management, 81% for economics & econometrics, and 84% for accounting & finance.

A possible compromise which incorporates elements of both of these adjustment methods is used here. It is calculated as follows: (i) outputs with missing ABS scores are given a score of unity if they are *neither* an authored book *nor* a journal article (e.g. edited books, book chapters, internet publications or working papers); (ii) authored books and journal articles are given a score of 2 if an ABS score is available *for only one output*; (iii) authored books and journal articles are given a score equal to the mean score of the outputs for which an ABS score is available if there are *at least 2 outputs* for which an ABS score is available. It turns out that all three methods of amending the ABS score are very highly correlated with each other. Furthermore, the main findings of the regression analysis are insensitive to whichever of the three amended ABS scores is used. Since the compromise method of adjusting the ABS score seems the most plausible *a priori*, this is used in the empirical analysis reported in section V below.

The three ABS scores per Category A staff are given for all three units of assessment in Table 4. In addition, a distinction is made between three groups of staff in the business & management unit: those publishing in economics journals, those publishing in accounting and finance journals, and the rest. This allows for a comparison of the mean ABS scores between different sub-groups. The overall impression is that the differences in the journal quality scores between the three units of assessment are consistent with the research output profiles given in Table 2. For example, the economics & econometrics unit of assessment has a significantly greater score than the other two units for all three measures of the ABS score (at the 5% level or less). We also note that when researchers in accounting & finance are split between those in business schools and those in autonomous accounting & finance departments, the mean ABS score is invariably larger for accounting & finance staff located in business schools than for any of the other staff groups across all four measures of the ABS score. A different result is obtained for economists: the ABS score for economists in business schools is not significantly different (at the 5% level) from the ABS score obtained by economists in autonomous economics departments. The claim that the outputs in economics & econometrics were of higher quality in autonomous economics departments than those cross-referenced from business & management (HEFCE 2009b) is not therefore supported by the ABS scores given in Table 4.

A second factor which may influence the productivity of researchers is the presence of scale economies (Johnes, Taylor and Francis, 1993; Taylor, 1994, 1995) resulting from having a large number of active researchers. Larger departments may offer advantages in the form of access to more expertise from colleagues working in close proximity to each other, especially if this also engenders greater competition between fellow researchers. But if departmental size affects the quality of research, this should be captured by the journal quality score. There is no obvious reason why departmental size should have an independent impact on research quality, unless the panel's judgment is biased in favour of large departments due to their visibility. We return to this problem in section V below.

The panel's judgment about a unit's research quality may also be influenced by the international reputation and status of the institution to which it belongs. In other words, individual departments may benefit from a 'halo' effect independently of the quality of its research output due to the presence of implicit bias in organizational decision-making (Nisbett and Wilson, 1977; Wilson *et al*., 1993; Reich *et al*., 2007; Jost *et al*., 2009). In the context of the RAE, it is possible that panel members were unduly influenced by the long-standing reputation of certain institutions in producing high quality research, even though panel members may have been determined not to be influenced by long-established reputational effects. Though these reputational effects are difficult to measure directly, it is nevertheless possible to distinguish institutions by their membership of well-defined clubs, such as the Russell Group of twenty universities, which was established in 1994 and includes several institutions with a long-established international reputation, such as Oxford, Cambridge, LSE and UCL.

A further group of institutions aspiring to be recognized for their research excellence is the 1994 Group of 'research-intensive' universities, the formation of which was an immediate response to the establishment of the Russell Group. Binary variables are therefore included to identify universities which are members of these two 'research-led' groups. These binary dummies are also included as explanatory variables in the esteem and research reputation regressions discussed in the results section below.

Esteem

The esteem bestowed on individual departments is inevitably subjective and is likely to be correlated with a department's research performance over several previous years. Since the vast majority of departments have been evaluated in previous RAEs, and since the panel's judgment of a department's esteem is likely to be influenced by its performance in previous RAEs, a department's outcome in the 2001 RAE can be used to explain its esteem in 2008.

Since highly-rated departments receive greater financial rewards than less successful departments, a successful outcome in the RAE may result in the subsequent expansion of research staff. Over time, the most highly rated departments may therefore be expected to grow more rapidly than less successful departments. Moreover, they will also be in a better position to attract high quality research staff, thereby resulting in a process of cumulative causation as expansion raises the average quality of research output, leading to still greater esteem. The size of a department may therefore be expected to be positively related to esteem due to a process of two-way causation.

The panel's judgment over a department's esteem may also be fueled by the halo effect bestowed on high-ranking institutions, as discussed above. Separate dummies for Russell Group and 1994 Group members are therefore included in the model.

Research environment

The RAE panel's view of the quality of a department's research environment is likely to be influenced by two main factors: the number of research students (especially those working towards a doctorate) and the amount of research income brought in (especially from the Research Councils). Table 5 shows the number of research students and the value of research grants expressed as a proportion of a department's research staff. There are wide variations in the amount of research income between the three units of assessment within some of the categories, such as Research Council income, this being much higher for economics & econometrics than for the other two units of assessment. Research income per staff is especially low for accounting & finance. Differences in research students per staff are less divergent between the three units. The number of research staff may also affect the quality of the research environment in so far as there may be a critical mass affect, in which case a non-linear relationship may be expected.

Finally, it is possible that panel members were again influenced by a halo effect in addition to the quantitative indicators based on research grants and research students. The binary variables indicating membership of the Russell Group and the 1994 Group are therefore included in the model.

Equations to be estimated

The foregoing discussion of the factors likely to influence the panel's judgment regarding the quality of a department's research output, esteem and research environment suggest the following three estimating equations:

$$\text{Research output}_i = \alpha_0 + \alpha_1\text{ABS\_score}_i + \alpha_2\text{Res\_staff}_i + \alpha_3\text{Russell}_i + \alpha_4\text{Group94}_i + \varepsilon_{1i} \qquad (1)$$

$$\text{Esteem}_i = \beta_0 + \beta_1\text{RAE01}_i + \beta_2\text{Res\_staff}_i + \beta_3\text{Russell}_i + \beta_4\,\text{Group94}_i + \varepsilon_{2i} \qquad (2)$$

$$\text{Research environment}_i = \gamma_0 + \gamma_1\text{Res\_inc}_i + \gamma_2\text{Res\_studs}_i + \gamma_3\text{Res\_staff}_i + \gamma_4\text{Res\_staff}_i^2 + \gamma_5\text{Russell}_i + \gamma_6\,\text{Group94}_i + \varepsilon_{3i} \qquad (3)$$

where:

Research output = RAE panel's assessment of each department's research output (see notes to Table 2)

Esteem = RAE panel's assessment of each department's esteem

Research environment = RAE panel's assessment of each department's research environment

Journal rating score = ABS journal rating score (see Table 4 and main text for details of its construction)

Res_staff = Number of Category A staff submitted to the RAE

RAE_2001 = RAE rating in 2001

Res_inc = income from Research Councils

Res_studs = number of research students

Russell = member of Russell Group of universities (=1 and zero otherwise)

Group94 = member of 1994 Group of universities (=1 and zero otherwise)

i = department in university i

Since the three dependent variables are the consequence of decisions reached by the same panel of experts, the equations are jointly estimated using a seemingly-unrelated regression model (SUR) in order to allow for correlated errors between the three equations. A problem with SUR estimation is that it eliminates any observations not common to all equations and its application leads to a small reduction in the number of observations in the present statistical analysis. This is due to missing values for the 2001 RAE ratings for those departments not assessed in both 2001 and 2008. The equations were therefore also estimated for the full sample using ordinary least squares (OLS) and weighted least squares (WLS) to check that the results were robust to the method of estimation and to the inclusion of the missing observations. The results were substantively unchanged.

V RESULTS

The three equations specified in section V are estimated for two units of assessment, namely, business & management and economics & econometrics. The number of observations is too small (i.e. 14) in accounting and finance for reliable estimates to be obtained. An equation has been estimated, however, for a combined group including all three units of assessment. Two binary variables are therefore added to the model in the combined equation, one identifying autonomous economics & econometrics departments and the other identifying autonomous accounting & finance departments. This allows us to investigate whether there are any differences between the three panels in their assessments (for each of the three dependent variables) *after controlling for the other explanatory variables*. The results for the seemingly-unrelated regressions are given in Table 6. The three equations specified in section V were also estimated using OLS and WLS. We note at the outset that the results are very similar regardless of the regression method used.[11]

Research output

Two results stand out in the research output regressions. First, the estimated coefficient on the journal quality score is highly significant and this variable alone explains a large proportion of the variation between departments in the panel's evaluation of research quality. This is the case for both business & management and for economics & econometrics as well as for the regression for all three units of assessment combined. The second main result is the highly significant coefficient on the size of departments, indicating that larger departments attained higher evaluations of research quality from the RAE panel than smaller departments.

A further result of interest is the highly significant coefficient on the dummy variable identifying autonomous economics departments in the combined regression. The estimated effect is substantial since it suggests that the economics & econometrics panel rated the research output of autonomous economics departments higher than business & management departments by a factor of 0.34 on the RAE rating scale for research output (which has a mean of 2.43 and a standard deviation of 0.45 over the 137 observations) *after controlling for the quality of publications*. It should also be noted that the Russell Group dummy is statistically significant (at 1%) in the combined equation, but the estimated impact of this halo effect is relatively small (at 0.11).

Esteem

[11] The OLS and WLS results are available in a working paper. We also note that the results are substantively unchanged when the regressions are estimated using the rank of each of the dependent variables and the rank of the ABS journal quality score rather than the actual values of these variables.
(See http://www.lums.lancs.ac.uk/publications/viewpdf/006168/)

The esteem of a department, as judged by the RAE panel, is significantly correlated with its 2001 RAE rating in all three regressions (at the 5% level or higher), but its estimated impact on esteem is nearly three times larger for business & management (at 0.25) than for economics & econometrics (0.09). Departmental size is also highly correlated with esteem in all three regressions. This result is consistent with the view that the RAE panel's judgments have been influenced by a department's visibility, which is partly determined by size. One problem with including departmental size as an explanatory variable, however, is that it may not be independent of a department's previous RAE outcome, since a successful outcome in the 2001 RAE may have been followed by an expansion as a consequence of the extra resources acquired. Such a two-way relationship will bias upwards the estimated impact of size on esteem.

The most interesting results obtained in the esteem regressions, however, relate to the dummy variables. The coefficient on the dummy variable identifying autonomous economics departments indicates that the economics & econometrics panel rated these departments 0.55 points higher than business & management departments. This is a substantial and highly significant disparity given an overall mean across the 137 departments of 2.43 and a standard deviation of 0.76. The reason behind this highly significant difference between the economics & econometrics panel and the business & management panel in their rating of a department's esteem is unclear since a department's previous RAE outcome is included specifically to capture this esteem effect on the 2008 outcome. A possible explanation is that the economics & econometrics panel were less demanding in their interpretation of esteem than the business & management panel, thus implying an element of upward bias in the economics & econometrics panel's judgment (or conversely a downward bias in the judgment of the business & management panel).[12]

The esteem score is also highly significantly related to the Russell Group dummy for both business & management and for economics & econometrics. Moreover, the estimated effect is of substantial magnitude, indicating a hike of around 0.60 in the esteem score for the Russell Group and a hike of 0.46 for the 1994 Group. Once again, these results suggest that the panels have been biased in favour of these two groups of universities in their evaluations of a department's esteem.

Research environment

There is very little evidence that the research environment score is related to either research income or to the number of research students. As in the assessment of research output and esteem, however, there is strong evidence that the panel's judgment has been influenced by a department's size, with larger departments attaining significantly higher scores for the quality of their research environment.

The most important explanatory variables, however, are the binary dummies. The economics & econometrics panel awarded autonomous economics departments a research environment score 0.63 higher than the business & management panel. This is an extremely large differential given the mean rating of 2.43 and the standard deviation of 0.68. This result is again consistent with the view that the economics & econometrics panel were less demanding than the business & management panel in their judgments. The

[12] Upward bias by the economics & econometrics panel seems more likely given that economics & econometrics had the highest overall RAE outcome score of all 67 units of assessment.

differential for accounting & finance was smaller (at 0.41) than that for economics & econometrics but still highly significant and substantial in magnitude. This suggests that attempts by the main panel to "ensure comparability of standards across its sub-panels" (HEFCE, 2009b) were unsuccessful in this case.

There is also evidence, once again, that the RAE panels have treated both the Russell Group and the 1994 Group of universities much more favourably than universities not in these two groups. The business & management panel, for example, rated the research environment of Russell Group and 1994 Group universities 0.51 and 0.47 higher, respectively, than other universities. The regression results for the research environment category of assessment are therefore consistent with the view that the judgment of the panels has been subject to bias.

Panel membership

Finally, since Doyle *et al.* (1995) found that departments which had a member on the panel achieved significantly higher outcomes in the 1992 RAE, a panel membership dummy was added to the regression equations. Each department with a staff member on the panel was identified by a binary variable. The estimated coefficient on the panel membership dummy was found to be very close to zero and was not statistically significant in all regressions (results not reported), thus supporting the null of zero bias across all three research profiles.

VI DISCUSSION

The claim that bibliometric data, such as journal quality ratings, offer a sound metric for determining the quality of research at departmental level is, at best, only partially supported by the statistical analysis of the 2008 RAE outcomes for the three cognate units of assessment investigated in this paper (viz. business & management, economics & econometrics and accounting & finance). The fact that there is a highly significant relationship between a journal rating score based on the ABS *Journal Quality Guide* and the overall RAE outcome for these three units of assessment is not sufficient evidence to support the proposal that metrics should play a predominant role in future assessments of research quality. This view ignores two important factors.

First, the regression equations reported here include other variables in addition to an index of journal quality, and most of these additional explanatory variables are inappropriate for assessing a department's research quality (though they may still be important in explaining the decisions reached by the RAE panel). One of the main findings in the regression analysis is the importance of a department's size in accounting for its research performance. This result is common to all three categories of research performance (research output, esteem and research environment). The justification for including 'size' as a regressor is that it may capture benefits arising from scale economies in so far as a critical mass of researchers in specific areas of activity may improve research productivity. Additionally, large departments may allow individual researchers to gain from a wider spread of research expertise than can be found in small departments. Larger departments may also find it easier to allocate more research time to staff as a result of scale economies in teaching and administration.

But if size does have a beneficial effect on research quality, this should be captured by the quality of the publications as indicated by the ABS score. It is not clear why size should have any *additional* effect on the quality of research output independent of the journal quality score. An alternative, and perhaps more plausible, explanation for the highly significant coefficient on departmental size is that the RAE panel regarded large departments as being more likely to produce high quality research. In other words, there may have been some implicit bias in the minds of panel members in favour of larger and hence more 'visible' departments. If this is the case, it would clearly be inappropriate to reward departments simply for their 'bigness'. There is no clear justification for discriminating against 'smallness' in the allocation of research funds.

The second critical problem with using the journal quality score to allocate research funds is that the high degree of explanatory power is partly the consequence of spurious correlation due to a small number of outliers. Omitting these outliers from the estimated regression equations, results in a significant reduction in explanatory power. For example, regressing the overall RAE outcome on the journal quality score for business & management, together with the number of Category A staff, we obtain an R^2 of 0.87. This falls to 0.72 when departmental size is dropped from the model and when just five of the ninety observations are omitted (i.e. those departments with an overall research outcome score of less than 1.4 or greater than 3.2). The corresponding reduction in R^2 for economics & econometrics is from 0.89 to 0.66 when size and just two outliers (with very low research outcome scores) are omitted.

The problem with using the journal quality score for allocating research funds can be demonstrated by comparing the *actual* RAE outcome with the outcome *predicted* from the journal quality score. A more interesting alternative is to convert the actual and predicted RAE outcomes into funding per staff, which can be done from information provided by HEFCE on the amount of funding allocated to each department as a consequence of the RAE. The scatter diagrams shown in Figures 2 and 3 show the actual £QR (i.e. block research grant) funding per staff plotted against each department's £QR funding per staff *predicted from its overall research performance* in the 2008 RAE. The predicted £QR funding per staff is obtained as follows. First, the overall RAE outcome for each department is regressed on each department's journal quality score. Second, this estimated relationship is then used to obtain the predicted RAE outcome for each department. Third, £QR funding per staff is regressed on the overall RAE outcome to obtain an equation for transforming the *predicted* RAE outcome into a *predicted* £QR funding per staff. Figure 2 plots the actual £QR funding per Category A staff in business & management departments against its predicted value (based on each department's journal quality score). Figure 3 provides the same information for economics & econometrics.

Thus, for business & management, we see that Keele and Swansea have virtually the same *actual* £QR funding per staff but that Swansea's *predicted* £QR funding per staff is more than twice that of Keele (see Figure 2). Similarly, both Durham and Queen Mary have a *predicted* £QR funding per staff nearly 25% greater than Cambridge, yet their *actual* funding per staff is 31% less. Likewise in economics &

econometrics, Royal Holloway has the highest *predicted* £QR funding per staff (based on its journal quality score), yet its *actual* £QR funding per staff is 33% below that of UCL and the LSE. On the assumption that the RAE panels 'got it right', these examples aptly demonstrate the need for considerable caution in using the journal quality score in the research assessment process. On the contrary, there is considerable discrepancy between the ranking of departments based on the RAE outcome and the ranking based on the journal quality score. This raises the question as to whether and to what extent the journal quality score should be used to aid the panel's judgments.

VII CONCLUSION

The primary aim of this paper has been to investigate the role that quantitative indicators should play in the research assessment process, drawing on the experience of the UK's 2008 Research Assessment Exercise. This is important because the research assessment process determines the allocation of research funds between universities. Regression methods have been used to explore the statistical relationship between the outcomes of the 2008 RAE and a set of quantitative indicators, including a journal quality score. The statistical analysis was undertaken on data covering three cognate units of assessment: business & management, economics & econometrics, and accounting & finance.

The four main findings are as follows:

i. A highly significant statistical relationship exists between *each of the three components* of research quality (i.e. research output, esteem and research environment) and various quantitative indicators, such as a journal quality score, departmental size and previous research assessment outcomes.

ii. The explanatory power of the estimated regression equations falls considerably when departmental size is omitted from the model and when a handful of outliers are also excluded.

iii. The inclusion of dummy variables indicates that departments which are members of the Russell Group or the 1994 Group of universities experienced substantially higher research quality scores than non-members, even after controlling for other variables such as the journal quality score and size. The reasons for this are not known, but it seems likely that these institutions benefit from a halo effect based on their reputation over the longer-term. If this is the case, the implication is that there has been bias in the decisions reached by both the business & management panel and the economics & econometrics panel in determining the overall research outcomes.

iv. There is evidence that the economics & econometrics panel was more generous in its assessment of research quality than the business & management panel. This suggests an element of bias in the decisions reached by the economics & econometrics panel relative to the business & management panel.

These findings offer some support for the use of journal quality indices in the research assessment process in so far as there is a high correlation between measures of journal quality and the judgments reached by RAE panels. But there is insufficient evidence to support the proposal that journal quality indices are sufficient *per se* for determining the allocation of research funds between institutions. This does not mean, however, that we should rely entirely on peer review since there is evidence that the RAE panels may have been significantly biased in their judgments. This would certainly be consistent with the psychology literature investigating the presence of implicit bias. Requiring the RAE panels to take journal quality scores into account, in addition to their more detailed analysis of individual research outputs, may help to reduce such bias in the future. Peer review supported by bibliometric data would seem to be the most favourable option for future assessments.

REFERENCES

Adler, N. and A.-W. Harzing (2009) 'When knowledge wins: Transcending the sense and nonsense of academic rankings', *Academy of Management Learning and Education*, **8**, 72-95.

Baruch, Y. and Hall, D. T. (2004) 'The academic career: A model for future careers in other sectors?', *Journal of Vocational Behavior*, **64**, 241-262.

Bence, V. and Oppenheim, C. (2004) 'The influence of peer review on the research assessment exercise', *Journal of Information Science*, **30**, 347-368.

Bessant, J., Cooper, C., Dawson, S., Gennard, J., Gardiner, M., Gray, A., Jones, P., Mayer, C., McGee, J., Pidd, M., Rowley, G., Saunders, J. and Stark, A. (2003) 'The state of the field in UK management research: reflections of the Research Assessment Exercise Panel', *British Journal of Management*, **14**, 51-68.

Butler, L. and McAllister, I. (2009) 'Metrics or peer review? Evaluating the 2001 UK Research Assessment Exercise in political science', *Political Studies Review*, **7**, 3-17.

Clark, T. and M. Wright (2007) 'Reviewing journal rankings and revisiting peer reviews: editorial perspectives', *Journal of Management Studies*, **44**, 612-621.

Cooper, C. and Otley, D. (1998) 'The 1996 Research Assessment Exercise for business and management', *British Journal of Management*, **9**,73-89.

Doyle, J.R., Arthurs, A.J., Green, R.H., Mcaulay, L., Pitt, M.R., Bottomley, P.A. and Evans, W. (1995) 'The judge, the model of the judge, and the model of the judged as judge: Analyses of the UK 1992 Research Assessment Exercise data for Business and Management Studies', *Omega Journal of International Management Science*, **24**, 13-28.

Easton, G. And Easton, D. M. (2003) 'Marketing journals and the Research Assessment Exercise', *Journal of Marketing Management*, **19**, 5-24.

Fleet, D. D. van, McWilliams, A. And Siegel, D. S. (2000) 'A theoretical and empirical analysis of journal rankings: the case of formal lists', *Journal of Management*, **26**, 839-861.

Geary, J., Marriott, L. and Rowlinson, M. (2004) 'Journal rankings in business and management and the Research Assessment Exercise in the UK', *British Journal of Management*, **15**, 95-141.

HEFCE (2003) *Joint consultation on the review of research assessment*, Consultation by the UK funding bodies on the review by Sir Gareth Roberts, May 2003/22, Higher Education Funding Council, Bristol. http://www.hefce.ac.uk/pubs/hefce/2003/03_22.htm

HEFCE (2009a) *Interim report of the REF bibliometrics pilot exercise*, June, Higher Education Funding Council, Bristol.

HEFCE (2009b) Subject overview report, Main Panel I, RAE2008, Higher Education Funding Council, Bristol.

Hobbs, F. D. R. and Stewart, P. M (2006) 'How should we rate research?' *British Medical Journal*, **332**, 983-984.

House of Commons (2004a) *Research Assessment Exercise: a re-assessment*, Science and Technology Committee, Eleventh Report of Session 2003–04, HC 586, September, London. http://www.parliament.the-stationery-office.co.uk/pa/cm200304/cmselect/cmsctech/586/586.pdf

House of Commons (2004b) *Research Assessment Exercise: a re- assessment: Government Response to the Committee's Eleventh Report of Session 2003–04*, First Special Report of Session 2004–05, Science and Technology Committee, HC 34, November, London. http://www.parliament.the-stationery-office.co.uk/pa/cm200405/cmselect/cmsctech/34/34.pdf

HM Treasury (2006a) *Science and innovation investment framework 2004-2014: next steps*, Budget, March 2006, London. http://www.hm-treasury.gov.uk/bud_bud06_adscience.htm

HM Treasury (2006b) *Investing in Britain's potential: Building our long-term future*, Pre-Budget Report 2006, Cm 6984, London. http://www.hm-treasury.gov.uk/d/pbr06_completereport_1439.pdf

Johnes, J., Taylor, J. and Francis, B. (1993) 'The research performance of UK universities: a statistical analysis of the results of the 1989 Research Selectivity Exercise', *Journal of the Royal Statistical Society* A, **156**, 271-286.

Johnston, R. (2009) 'Where there are data: Quantifying the unquantifiable', *Political Studies Review*, **7**, 50-62.

Jost. J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J. and Hardin, C.D. (2009) 'The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore', *Research in Organizational Behavior,* 29, 39–69.

Kalaitzidakis, P., Mamuneas, T. P. and Stengos, T. (2003) 'Rankings of academic journals and institutions in economics', *Journal of the European Economic Association*, **1**, 1346-1366.

Kelly, A., Morris, H., Rowlinson, M. and Harvey, C. (2009a) *Academic Journal Quality Guide, Version 3,* Association of Business Schools, March, London. http://www.the-abs.org.uk/files//ABS2009web_introduction.pdf

Kelly, A., Morris, H. and Harvey, C. (2009b) *Modelling the outcome of the UK Business and Management Studies RAE 2008 with reference to the ABS Journal Quality Guide*, June, Association of Business Schools, London. http://www.the-abs.org.uk/files//RAE2008_ABS2009_final.pdf

Laband, D. N. and Piette, M. J. (1994) 'The relative impacts of economics journals: 1970-1990', *Journal of Economic Literature*, **32**, 640-666.

Lee, F. S. (2007) 'The Research Assessment Exercise, the state and the dominance of mainstream economics in British universities', *Cambridge Journal of Economics*, **31**, 309-325.

Liebowitz, S.L. and Palmer, J. P. (1984) 'Assessing the relative impacts of economics journals', *Journal of Economic Literature*, **22**, 77-88.

Liner, G. H. (2002) Core Journals in Economics, *Economic Inquiry* 40(1), 138-145.

Macdonald, S. and J. Kam (2007) 'Ring a ring o'roses: quality journals and gamesmanship in management studies', *Journal of Management Studies,* **44**, 640-655.

McLean, I., Blais, A., Garand, J. C. and Giles, M. (2009) ‘Comparative journal ratings: A survey report’, *Political Studies Review*, **7**, 18-38.

Mingers, J., Watson, K. and Scaparra, P. (2009) ‘Estimating business and management journal quality from the 2008 Research Assessment Exercise in the UK’, Working Paper 205, Kent Business School.

Nierop, R. van (2009) ‘Why do statistics journals have low impact factors?’, *Statistica Neerlandica,* **63**, 52-62.

Nisbett, R. E. and Wilson, T. D. (1977) ‘The halo effect: Evidence for unconscious alteration of judgments’, *Journal of Personality and Social Psychology,* **35**, 250-256.

Nkomo, S. (2009)’The seductive power of academic journal rankings: Challenges of searching for the otherwise’, *Academy of Management Learning and Education*, **6**, 106-112.

Ozbilgin, M. F. (2009) ‘From journal rankings to making sense of the world’, *Academy of Management Learning and Education*, **6**, 113-121.

Otley, D. (2002) ‘British research in accounting and finance (1996-2000): the 2001 Research Assessment Exercise’, *British Accounting Review,* **34**, pp.387-417.

Paul, R. J. (2008) ‘Measuring research quality: The united Kingdom Government's Research Assessment Exercise’, *European Journal of Information Systems*, **17**, pp. 324-329.

Reich, D. A., Green, M. C., Brock, T. C. and Tetlock, P. E. (2007) ‘ Biases in research evaluation: inflated assessment, oversight or error-type weighting?’, *Journal of Experimental Psychology*, **43**, 633-640.

Roberts, G. (2003) *Review of Research Assessment: A Report to the Funding Bodies*, HEFCE report 2003/22, Bristol. http://www.ra-review.ac.uk/reports/.

Russell, A. (2009) ‘Retaining the peers: how peer review triumphs over league tables and faulty accounting in the assessment of political science research’, *Political Studies Review*, **7**, 63-72.

Seglen, P. O. (1997)’Why the impact factor of journals should not be used for evaluating research’, *British Medical Journal*, **314**, 498-502.

Singh, G., Haddad, K.M. and Chow, C. W. (2007) ‘Are articles in "top" management journals necessarily of higher quality?’ *Journal of Management Inquiry*, **16**, 319-331.

Starbuck, W. H. (2005) ‘How much better are the most-prestigious journals? The statistics of academic publications’, *Organization Studies*, **16**, 180-200.

Taylor, J. (1994) ‘Measuring research performance in business and management studies in the United Kingdom: the 1992 Research Assessment Exercise’, *British Journal of Management*, **5**, pp.275-288.

Taylor, J. (1995) ‘A statistical analysis of the 1992 Research Assessment Exercise’, *Journal of the Royal Statistical Society A*, **158**, 241-261.

Taylor, J. and Izadi, H. (1996) ‘The 1992 Research Assessment Exercise: outcome, outputs and inputs in economics and econometrics’, *Bulletin of Economic Research*, 48, pp. 1-26.

Tetlock, P. E. and Mitchell, G. (2009) ‘Implicit bias and accountability systems: What must organizations do to prevent discrimination?’ *Research in Organizational Behavior*, 29, 3-38.

Tunzelmann, N. von, Mbula, E. K. (2003) *Changes in research assessment practices in other countries since 1999: final report*, February, Science Policy Research Unit, University of Sussex.

http://www.ra-review.ac.uk/reports/Prac/ChangingPractices.pdf

Wall, H. (2009) 'Journal rankings in economics: handle with care', Working Paper 2009-014A, Federal Bank of St. Louis. http://research.stlouisfed.org/wp/2009/2009-014.pdf

Weale, A. R., Bailey, M. and Lear, P. A. (2004) 'The number of non-citation articles within a journal as a measure of quality: a comparison to the impact factor', *BMC Medical Research Methodology*, **4**, 14.

Williams, G. (2007) 'Should we ditch impact factors?', *British Medical Journal*, 334, 568.

Wilson, T. D., DePaulo, B. M., Monk, D. G. and Klaaren, K. J. (1993) Scientists evaluation of research: the biasing effects of the importance of the topic', *American Psychology Association*, **4**, 322-325.

TABLE 1 Percent of designated staff in each research quality category (weighted by the number of Category A staff in each assessed unit)

Unit of assessment	Overall research quality profile (weighted): % of staff in each category				
	4*	3*	2*	1*	Unclassified
Research output (weight=0.7)					
Business & Management	13.8	36.2	35.4	13.5	1.0
Economics & Econometrics	26.5	47.2	24.1	2.1	0.1
Accounting & Finance	6.0	36.4	40.7	15.9	1.0
Esteem (weight=0.2)					
Business & Management	26.7	34.5	25.3	10.7	2.8
Economics & Econometrics	20.4	68.1	11.3	0.3	0.0
Accounting & Finance	18.9	35.3	33.1	12.7	0.0
Research environment (weight=0.1)					
Business & Management	23.0	37.4	27.9	10.1	1.6
Economics & Econometrics	30.1	49.1	20.4	0.3	0.0
Accounting & Finance	7.3	41.9	34.4	14.1	2.3
Overall RAE classification					
Business & Management	16.6	36.6	32.9	12.8	1.0
Economics & Econometrics	26.6	49.3	22.9	1.2	0.0
Accounting & Finance	7.7	36.6	39.3	15.0	1.3

Notes:
4* = world-leading in terms of originality, significance and rigour
3* = internationally excellent in terms of originality, significance and rigour but falls short of the highest standards of excellence
2* = internationally recognised in terms of originality, significance and rigour
1* = nationally recognised in terms of originality, significance and rigour
Unclassified = falls below the standard of nationally recognised work
Source: RAE2008 (http://www.rae.ac.uk/results/).

TABLE 2 Mean RAE score by quality profile category
(weighted by number of Category A staff in each assessed unit)

Unit of assessment	Mean scores in each quality profile category			
	Research output	Esteem	Research environment	Overall score
Business & Management	2.48	2.72	2.70	2.55
Economics & Econometrics	2.98	3.09	3.09	3.01
Accounting & Finance	2.31	2.60	2.38	2.34

Notes: The scores for the three research profiles (research output, research esteem and research environment) were calculated as follows:

i. Each of the five quality profile categories is given a score ranging from 0 to 4 (unclassified=0, 1*=1, 2*=2, 3*=3 and 4*=4).

ii. The score in each of the five categories is then multiplied by the proportion of research staff attributed to each category by the panel and summed to obtain the weighted mean for the specific quality profile.

iii. The overall RAE score (which is a combination of the three quality profiles) is obtained by using weights of 0.7, 0.2 and 0.1 for research output, research esteem and research environment respectively (as used in the RAE).

Source: Author's own calculations from RAE2008 database. See http://www.rae.ac.uk/results/.

TABLE 3 Research output submitted to the 2008 RAE by major publication category

Type of output	Business & Management	Economics & Econometrics	Accounting & Finance
Journal articles	90.5	84.9	86.1
Internet publications	2.5	9.5	6.0
Authored books	2.3	0.7	7.8
Other	4.7	4.7	5.1

Notes: 'Internet publications' include papers accepted for publication but still pending.

Source: Author's own calculations from RAE2008 database. See http://www.rae.ac.uk/results/.

TABLE 4 Journal quality score per Category A staff

Unit of assessment	Number of staff	Actual ABS score	Mean ABS journal quality score per category A staff: Amended ABS score I	Amended ABS score II	Amended ABS score III
Business & Management	3540	8.15	8.80	9.58	9.33
Economics & Econometrics	877	8.79	9.43	10.39	9.78
Accounting & Finance	168	7.61	8.14	8.69	8.47
Groups in Business & Management					
All other staff in business & management	2776	7.85	8.59	9.47	9.20
Economics staff in business & management	351	9.00	9.30	9.72	9.60
Accounting & finance staff in business & management	413	9.50	9.76	10.20	9.94

Notes:

1. The three groups in business & management are delineated according to whether articles were published in economics, accountancy or finance journals. Staff in business & management who published at least one-half of their articles in economics journals were attributed to the economics sub-group. Hence, the 32 staff who published one-half of their articles in economics journals and one-half in finance journals were attributed to the economics sub-group.
2. The ABS score is based upon a division of journals into four groups, each being given a score of between 1 and 4. A journal in the bottom group is given a score of 1 and a journal in the top group is given a score of 4. An ABS journal score is available for 13226 (82%) of the 16165 publications submitted by all three units of assessment.
3. Three amended ABS scores are estimated for each individual Category A staff by attributing a score to the missing values. Amended ABS score I grades all outputs for which an ABS score is not available as 1* (and therefore have a weight of unity). Amended ABS score II grades all outputs for which an ABS score is not available as equal to the mean of the non-missing values. Amended ABS score III was obtained as follows: (i) outputs with missing ABS scores were given a score of unity if they were neither an authored book nor a journal article (e.g. edited books, book chapters, internet publications or working papers); (ii) authored books and journal articles were given a score of 2 if an ABS score was available for only one output; (iii) authored books and journal articles were given a score equal to the mean score of the outputs for which an ABS score was available if there were at least 2 outputs for which an ABS score was available.

Source: Author's calculations from Association of Business Schools, *Journal Quality Guide*, 2009 and research outputs of all individual staff in RAE2008 (http://www.rae.ac.uk/results/).

Table 5 Research income and research students per Category A staff (weighted by number of Category A staff in each assessed unit)

	Business & management	Economics & econometrics	Accounting & finance
Research income (in £1000s) per Category A staff			
Research councils	26.7	58.0	3.0
Charities	6.2	12.1	3.9
Government	34.7	13.5	1.1
Industry	16.9	3.1	1.9
EU and other	22.1	17.0	6.7
Total	106.7	103.7	16.6
Research students per Category A staff			
Research students	2.1	1.7	1.2
PhDs awarded	1.0	0.6	0.7

Source: Author's own calculations from RAE2008 (http://www.rae.ac.uk/results/).

TABLE 6 Seemingly-unrelated regressions for research output, esteem and research environment

(a) Research output

Explanatory variables	Business & Management	Economics & Econometrics	B&M, Economics, Acc./Finance
Journal quality score	**0.772*****	**0.678*****	**0.733*****
	(0.047)	(0.086)	(0.042)
Number of Category A staff	**0.0021*****	**0.0057****	**0.0024*****
	(0.0005)	(0.0022)	(0.0005)
Economics & Econometrics dummy			**0.344*****
			(0.035)
Accounting & Finance dummy			0.050
			(0.048)
Russell Group	**0.076#**	0.098	**0.107****
	(0.044)	(0.092)	(0.038)
Group 94	0.071	-0.012	0.049
	(0.046)	(0.074)	(0.037)
Constant	0.285	0.807	0.365
	(0.106)	(0.218)	(0.095)
R^2	0.90	0.86	0.91
n	81	30	121

(b) Esteem

Explanatory variables	Business & Management	Economics & Econometrics	B&M, Economics, Acc./Finance
RAE rating 2001	**0.250*****	**0.092***	**0.130***
	(0.072)	(0.046)	(0.051)
Number of staff	**0.0085*****	**0.0079***	**0.0111*****
	(0.0023)	(0.0037)	(0.0019)
Economics & Econometrics dummy			**0.550*****
			(0.107)
Accounting & Finance dummy			0.263
			(0.168)
Russell Group	**0.560*****	**0.466*****	**0.596*****
	(0.142)	(0.133)	(0.109)
Group 94	**0.487*****	**0.213#**	**0.457*****
	(0.145)	(0.112)	(0.105)
Constant	0.862	2.119	1.183
	(0.195)	(0.166)	(0.144)
R^2	0.73	0.71	0.72
n	81	30	121

(c) Research environment

Explanatory variables	Business & Management	Economics & Econometrics	B&M, Economics, Acc./Finance
Research Council income (£m)	**0.081***	0.019	0.026
	(0.034)	(0.014)	(0.017)
Research grants from charities (£m)	0.065	0.087	0.095
	(0.122)	(0.059)	(0.068)
Research students	0.0005	-0.0001	0.0003
	(0.0003)	(0.0004)	(0.0003)
Number of Category A staff	**0.023*****	**0.023***	**0.019*****
	(0.003)	(0.010)	(0.003)
Number of Category A staff squared	-0.00012***	-0.0002	-0.00008***
	(0.00002)	(0.0001)	(0.00002)
Economics & Econometrics dummy			**0.634*****
			(0.085)
Accounting & Finance dummy			**0.412*****
			(0.120)
Russell Group	**0.511*****	**0.305***	**0.460*****
	(0.104)	(0.147)	(0.089)
Group 94	**0.473*****	0.187	**0.401*****
	(0.109)	(0.118)	(0.087)
Constant	1.345	2.297	1.471
	(0.082)	(0.132)	(0.078)
R^2	0.81	0.71	0.77
n	81	30	121

Notes: ( ) = standard errors; #, *, **, *** refer to significance at the 10%, 5%, 1% and 0.1% levels respectively. The computation of the covariance matrix for the equation residuals uses the small-sample adjustment procedure available in STATA. The Breusch-Pagan test rejects the null that the errors are uncorrelated for all three equations. The OLS and WLS results are not substantively different to the SURE estimates provided in this table (see working paper).

Binary variables were included in the regressions to identify whether or not a department belonged to a Russell Group or a 1994 Group university. The Russell Group was established in 1994 and comprises the following universities: Cardiff, ICL, KCL, LSE, Queen's Belfast, UCL, Birmingham, Bristol, Cambridge, Edinburgh, Glasgow, Leeds, Liverpool, Manchester, Newcastle, Nottingham, Oxford, Sheffield, Southampton and Warwick. The 1994Groupwas established in 1994 and comprises the following universities: Birkbeck College, Lancaster, Loughborough, Queen Mary, Royal Holloway, SOAS, Bath, Durham, East Anglia, Essex, Exeter, Leicester, Reading, St Andrews, Surrey, Sussex and York.

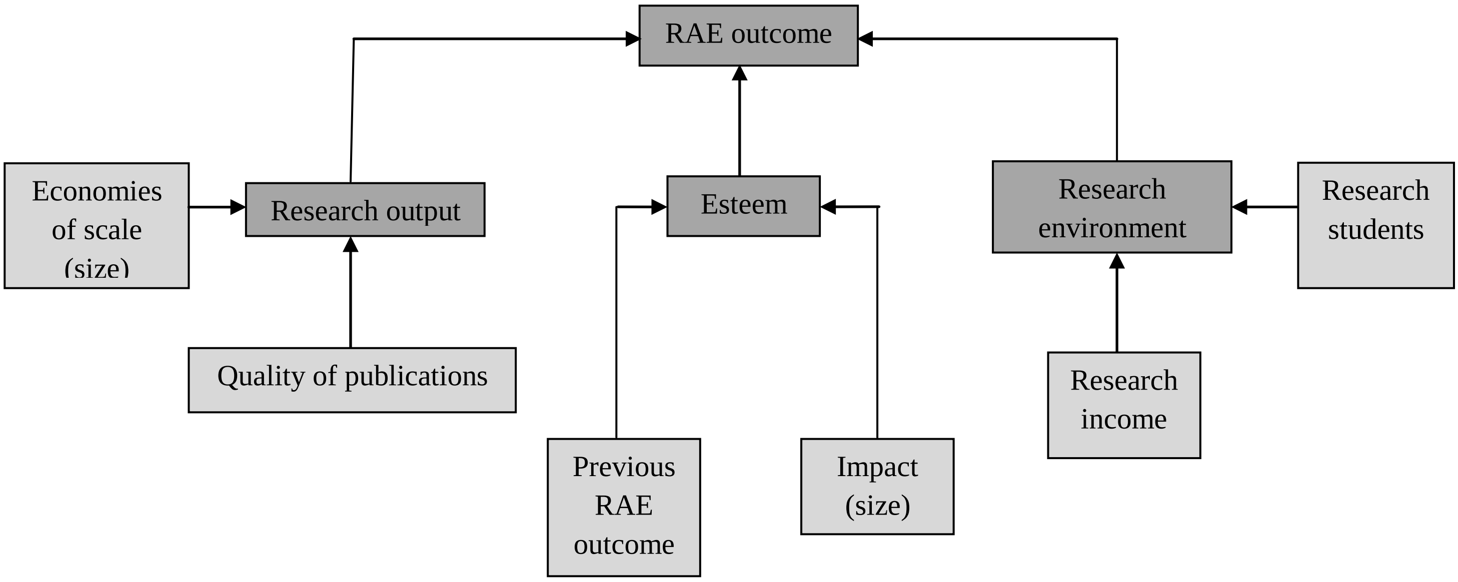
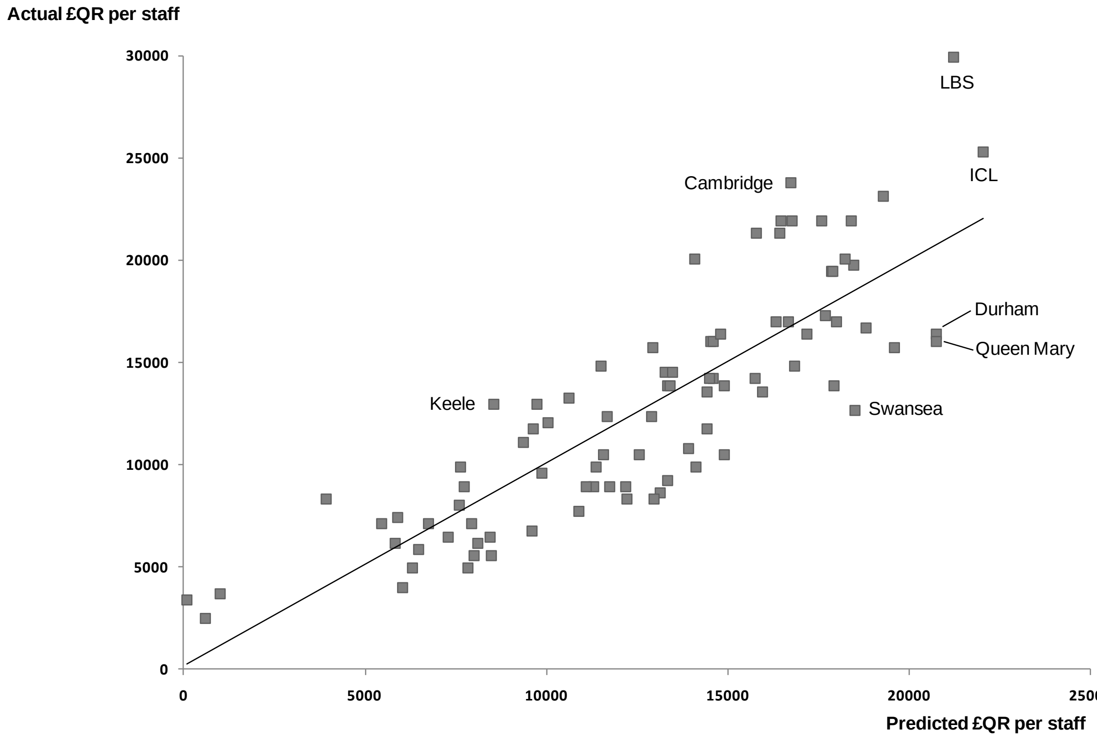
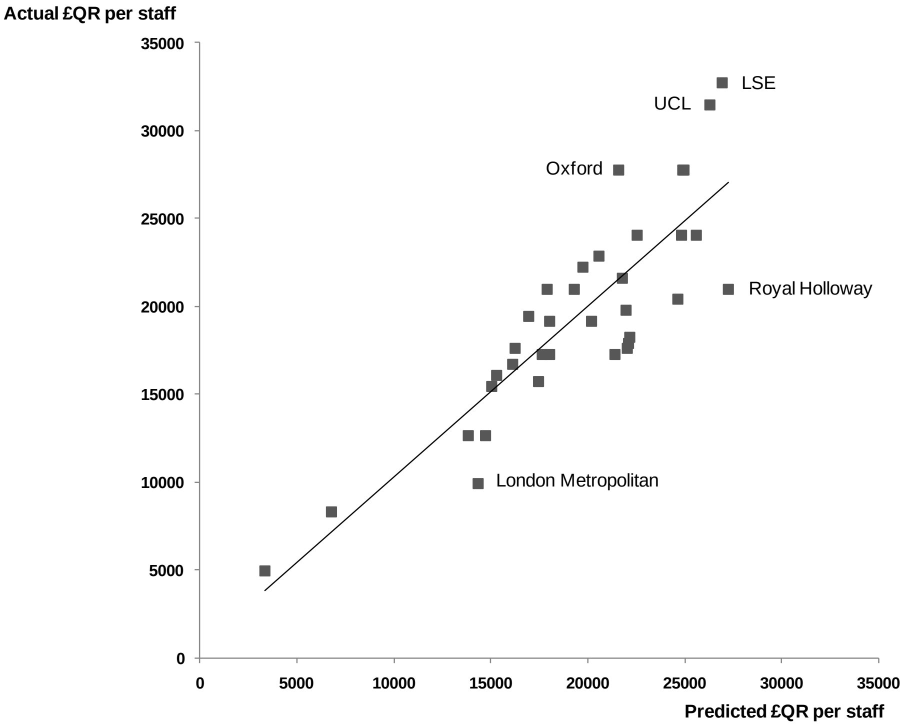


Figure 1 Determinants of the RAE outcome: research output, esteem and research environment



Notes: £QR per staff is the research funding calculated for each department by HEFCE for 2009/10 based on each department's overall research assessment score in the 2008 RAE and the number of Category A staff submitted for assessment. The predicted £QR per staff was obtained as follows: (i) the overall RAE outcome for each department was regressed on the journal quality score; (ii) the £QR per staff funding was regressed on the overall RAE outcome; (iii) the *predicted* £QR per staff was obtained by substituting the *predicted* overall RAE outcome, obtained from step (i), into the regression equation estimated in step (ii).
The £QR data were obtained from the HEFCE website: http://www.hefce.ac.uk/Research/Funding/qrfunding/.

Figure 2 Actual v predicted £QR allocation per staff: business & management



Notes: See notes to Figure 2.

Figure 3 Actual v predicted £QR allocation per staff: economics & econometrics